LEVEL II

Research Memorandum 78-23

GENERALIZED PACKAGES FOR ANALYSIS OF VARIANCE AND CATEGORICAL DATA

Roland J. Hart

AD A077970

ARI FIELD UNIT AT PRESIDIO OF MONTEREY, CALIFORNIA

DDC FILE COPY

DDC
RECEIVED
DEC 12 1979
D

U. S. Army
Research Institute for the Behavioral and Social Sciences

September 1978

DISTRIBUTION STATEMENT A
Approved for public release;
Distribution Unlimited

79 22 5 124

Army Project Number
16 2Q763744A769

Army Contemporary
Issues Development

9 memos
Research Memorandum 78-23

6 GENERALIZED PACKAGES FOR ANALYSIS OF VARIANCE AND CATEGORICAL DATA

12 27

10 Roland J. Hart
James A. Thomas, Work Unit Leader

ARI FIELD UNIT AT PRESIDIO OF MONTEREY, CALIFORNIA

11 September 1978

14 ARI-RM-78-23

Submitted as complete and
technically accurate by:
Jack J. Sternberg
Field Unit Chief

Approved by:

E. Ralph Dusek, Director
Individual Training and Performance
Research Laboratory

Joseph Zeidner
Technical Director
U.S. Army Research Institute for the
Behavioral and Social Sciences

Accession For
NTIS GRA&I [X]
DDC TAB [ ]
Unannounced
Justificatio...
By
Distribution/
Availability Codes
Dist. | Avail and/or special
A

Research Memorandums are informal reports on technical research problems. Limited distribution is made, primarily to personnel engaged in research for the Army Research Institute.

408 020

GENERALIZED PACKAGES FOR ANALYSIS OF VARIANCE AND CATEGORICAL DATA

This paper groups (a) analysis of variance and (b) categorical data problems into several classes and then describes general software packages that can analyze all classes of problems that have been defined. The strengths and weaknesses of a variety of software packages are compared in terms of the classes of problems they can handle and the ease with which they can be used. A method for analyzing unbalanced split-plot designs with currently available software is described.

ANALYSIS OF VARIANCE

Unbalanced Designs

Psychologists doing field research often have unequal sample sizes in different cells in analysis of variance designs. When unequal sample sizes exist, the design is considered unbalanced. There are some statistical complications associated with the least-squares analysis of unbalanced designs, and appropriate software is more difficult to find.

Statistical complications with unbalanced designs include several threats to validity of results. With balanced designs, the sums of squares that go into the numerators of the F ratios for each term in the model are independent. Independence does not exist in the term in the denominator, since the common sums of squares for error in the denominator are used to test a variety of terms in the model. With unbalanced designs, however, the term in neither the numerator nor the denominator of an F ratio is independent, which may create increased problems of Type I error, particularly in the case where many F tests are made on a large number of terms in an unbalanced design. With split-plot analysis of variance designs (where one or more factors are repeated-measures factors), F tests are approximate rather than exact, particularly when unbalance exists. The expected mean square coefficient for a term in the numerator of an F ratio will not be exactly the same as the coefficient for the same term in the denominator, in unbalanced split-plot designs, and the difference becomes larger as the unbalance becomes greater.

Another complication arises in the way the variance is partitioned with unbalanced designs. The researcher has the option of partitioning the variance in a variety of ways depending on his objectives. When unbalance exists, confounding between different terms in the model also exists. In other words, the expected mean squares for each term in the model contain a variety of extraneous components. These extraneous components include some or all of the expected mean square components for terms that come after the term of interest in the model statement.

The researcher's objective may be to construct F ratios that have the same expected mean square components entering into the mean squares as those found in the analogous balanced analysis of variance design--in other words, to unconfound expected mean squares for all terms in the model. In a fixed-effect factorial design, this would mean adjusting each term for all the other terms in the model (by ordering each term of interest last in the model statement). When this is done, the sums of squares for all terms in the model, when added together, are always less than the total sums of squares for the design. This approach involves assigning only that portion of the variance that is unconfounded to each term in the model and eliminating the confounded portion of the variance.

The researcher may not wish to partition the variance so that the expected mean squares are unconfounded by extraneous components. Instead, the researcher may wish to partition the variance in a hierarchical manner, so that the sums of squares for each term in the model, when added, equal the total sums of squares. This type of partitioning is done when the researcher is willing to assume for theoretical or practical reasons that some terms take precedence over others, e.g., main effects over interactions. When these assumptions are made, the expected mean squares for the terms that take precedence over others are still confounded with extraneous expected mean square components. However, making the assumption that some variables take precedence over others, and then partitioning the variance in a hierarchical manner consistent with these assumptions, is equivalent to taking that portion of the variance which is confounded and assuming that the confounded variance is due to the variable that takes precedence rather than to the variables that are confounded with it. This approach is a method of assigning unconfounded variance to the appropriate terms in the model, and then assigning that portion of the variance that is confound to one term rather than to other terms that are confounded with it, by making the assumption of precedence.

Multivariate Analysis of Variance

Psychologists doing field research face not only unbalanced designs but also designs with multiple dependent variables. A univariate analysis of variance is often computed for each of a large number of dependent variables, which creates the problem of inflation of Type I error. When multiple univariate F tests are made, some will be significant by chance alone. A problem with interpreting results arises when significance is found with univariate F tests at a level not much beyond what might be expected on the basis of chance.

Multivariate analysis of variance controls for this type of inflation of Type I error. Multivariate analysis of variance reduces each subject's scores on each of the dependent variables to one number, a number that is a simple linear combination of the subject's scores on

each of the original dependent variables. Multivariate analysis of variance consists of a search for the linear combination of dependent variables that discriminates best between levels of an independent variable in the sense of producing the largest possible univariate F ratio (Harris, 1975). Significance for this largest possible F ratio is determined by a critical value that is appropriate for it, one that takes into account the extreme capitalization on chance that was made in arriving at it. The original linear combination of dependent variables that discriminates best between levels of an independent variable is the same as the primary discriminant function found with discriminant analysis.

Since multivariate test statistics are based on a linear combination of dependent variables, there is no necessary one-to-one relationship between univariate F ratios and the multivariate test statistic. In other words, it is possible to have significant univariate F ratios but not to have significance with the multivariate test statistic, or the reverse--with a significant multivariate test statistic and no significant univariate F ratios. It is informative, however, if the researcher finds significance with the univariate F ratios but not with the multivariate test statistic. When this is the case, the researcher can best interpret the significant univariate F ratios as due to chance, i.e., the inflated Type I error that occurred when multiple F tests were made.

The nonsignificant multivariate test statistic indicates that it was not possible to find a linear combination of dependent variables that could produce a significant univariate F. In many cases, psychologists may wish to run multivariate analyses of variance to find out whether or not the significance that is found with multiple univariate analyses is real, i.e., whether it is due to inflated Type I error.

Each univariate analysis of variance design has a multivariate analogue. As mentioned previously, it is difficult to find appropriate software that can handle unbalanced analysis of variance designs and even more difficult for unbalanced multivariate analysis of variance designs. However, programs are available that can handle these designs.

Random and Mixed Effects

As mentioned previously, analysis of variance designs can be classified as (a) balanced or unbalanced and (b) univariate or multivariate. Other classifications of analysis of variance designs are also important: (c) the classification of the design as a random, fixed, or mixed model; and (d) the classification of the design as one with repeated-measures factors or one without such factors.

In the random model, the levels of the independent variables are randomly selected, and researcher wishes to generalize his results to all levels of the independent variable within the population of interest.

In the fixed model, the levels selected exhaust the population of interest, and the researcher wishes only to generalize to the selected levels that have been fixed. In the mixed model, some independent variables are fixed effects, and some random effects.

The classification of a design as a random, fixed, or mixed model affects the selection of the appropriate error term. With the random or mixed model, the expected mean squares for some or all terms in the model contain additional components that are not found in the fixed model. The appropriate error term must also contain the appropriate additional components in its expected mean square, to test the significance of a term that contains these additional components. With the random or mixed model it is sometimes necessary to construct error terms that contain the appropriate expected mean squares, including the appropriate additional components, by doing arithmetic on the mean squares of selected terms in the model and then making a quasi-F test with the constructed error term. It is also common to find that with random or mixed models, error terms with appropriate expected mean squares exist but contain too few degrees of freedom to make tests that are at all powerful.

Split-Plot Designs

The final classification of analyses of variance designs made here is between designs that contain repeated-measures factors and those that do not. Designs that have some factors with repeated measures and some factors without repeated measures are often called split-plot designs. In these designs, multiple measurements are made for the same subject or unit of analysis. These designs create a factor for between subjects (or units) differences, and the variance for this term is extracted from the error variance used to test the repeated-measures terms. Error variance is reduced in this manner, but so are the degrees of freedom for the error term(s) that are used to test the repeated terms. These models also require more restrictive assumptions of constant correlation between responses at all levels of each repeated-measures factor.

Software Comparisons

Analysis of variance software packages can be compared in terms of their generality--the extent to which they handle the categories of designs that were described previously: (a) balanced, unbalanced; (b) univariate, multivariate; (c) random, fixed; and (d) with or without repeated measures (split-plot or not).

The Biomedical package (BMD) for the most part deals only with balanced univariate analysis of variance designs (Dixon, 1973). The program BMD12V can handle balanced, multivariate, or univariate analyses of variance, including split-plot designs, for the fixed model only. BMD12V can only handle balanced designs. A limited number of unbalanced, univariate designs can be handled with other BMD programs. The programs assume for the most part a fixed-effects model. The Statistical Package for the Social Sciences (SPSS), release 06, has an analysis of variance program that will handle balanced or unbalanced univariate, fixed, factorial designs (Nie, Hull, Jenkins, Steinbrenner, & Bent, 1975). The program cannot handle multivariate analyses, random or mixed models, or split-plot designs.

The program has only three preestablished ways of partitioning the variance: (a) the "classic" option, in which main effects are adjusted for main effects only and interactions are adjusted for main effects and interactions; (b) the "regression" option, in which each term is adjusted for all other terms; and (c) the "hierarchical" option of adjusting each term only for the terms preceding it. With unbalanced designs, the regression option produces mean squares that are unconfounded by eliminating the confounded variance in the manner described previously. Unfortunately, there is a programing error with this option in release 06, which means it cannot be used. A multivariate analysis of variance program is planned for SPSS, release 07. The OSIRIS package contains a multivariate analysis of variance program, MANOVA, which can handle for the fixed model univariate or multivariate analyses of variance for balanced or unbalanced designs. This program cannot handle split-plot designs.

The most general analysis of variance programs that the author is aware of are (a) the MAD/RUMMAGE program (Bruce & Carter, 1974) and (b) the MULTIVARIANCE program (Finn, 1974).[1] MAD is the original and RUMMAGE is the updated version of the same generalized analysis of variance program. RUMMAGE is updated and improved on a continuing basis. Both MAD/RUMMAGE and MULTIVARIANCE can analyze any crossed or nested design, including split-plot designs, that are either balanced or unbalanced, for either univariate or multivariate analyses. These are the only programs that the author is aware of that can handle unbalanced

---

[1]The MAD/RUMMAGE package is available from Dr. Gale Bryce, Department of Statistics, 204 TMCB, Brigham Young University, Provo, Utah 84602, phone (801) 374-1211, extension 4505. Implementation is on IBM 360/370 as well as several non-IBM systems. Army Research Institute, Presidio of Monterey Field Unit, has a copy of MAD. The MULTIVARIANCE package is available from International Educational Services, P.O. Box A3650, Chicago, Illinois 60690, phone (312) 684-4920. Implementation is on IBM 360/370, CDC 6000 series, UNIVAC 1108.

split-plot designs. Both programs can do univariate and multivariate covariance analysis as well.

Design Statements. Both the MAD/RUMMAGE and the MULTIVARIANCE programs have their own strengths and weaknesses in particular areas, but together provide a powerful package that can handle practically any analysis of variance problem. The strengths and weaknesses of each program can be compared. One strength of the MAD/RUMMAGE program is the simplicity of the model statement that defines the analysis of variance design. The user enters model statements in a form familiar to most users, e.g.,

$$Y(IJ) = A(I) + B(J) + AB(IJ) + E.$$

If the user is not familiar with entering model statements in this form, he or she can quickly learn to write an analysis of variance model statement for any design by learning a few simple rules. For example, crossed and nesting relationships are indicated by the subscripts in parentheses. A nested relationship exists when there are more subscripts in parentheses associated with a given term than there are letters associated with that term. The "extra" subscripts define those terms that the term of interest is nested within. The user can quickly decide which interactions should be included and which ones should be excluded, when writing the complete full-rank model for any design, by referring to the following rule: Interactions between any two terms in the model should be included, if the subscripts that are in parentheses for the two terms that are being considered for interaction do not contain any subscripts that are the same. Interactions do not enter the model if two terms contain common subscripts in parentheses.

In contrast, the MULTIVARIANCE program requires the user to define the analysis of variance model by entering design matrices. It is more difficult to write correct design matrices, particularly with designs that include nesting and high order interactions, and more difficult to enter the multiple cards for these matrices into the program, than to write a single model statement as required by MAD/RUMMAGE.

Expected Mean Squares. A second strength of the MAD/RUMMAGE program is that it provides a matrix showing the expected mean squares for each term in the model. This is particularly useful when analyzing random or mixed model analysis of variance designs. In the balanced case, the user can immediately identify the appropriate error terms and identify how to construct error terms if appropriate ones do not exist. MULTIVARIANCE does not provide expected mean squares, so in this case the user would need to calculate them himself to find the appropriate error terms.

The expected mean square output from MAD/RUMMAGE helps the researcher identify confounding in unbalanced analyses of variance designs. The researcher's objective may be to arrive at unconfounded sums of squares that have the same expected mean squares as those found in an analogous balanced analysis of variance design. If the researcher is not sure about the way in which the terms in the model are confounded in an unbalanced design, the expected mean squares output from MAD/RUMMAGE will give the information, and the researcher can then order and reorder terms in the model to eliminate the confounding. In those cases where the researcher partitions the variance hierarchically, the expected mean squares output provides the researcher with information about the nature of the assumptions being made. The researcher is assigning confounded variance to one term rather than to others by assumption, and by identifying confounded terms, this assumption becomes explicit. The researcher can examine confounded terms to see if it is reasonable to assume that one confounded term takes precedence over the others.

Expected mean squares are also useful in identifying the confounding that occurs with incomplete block designs (i.e., designs in which there are missing cells). When there are missing cells (no observations in one or more cells), the resultant analysis requires the researcher to assume that particular interactions are zero in order to (a) eliminate confounding and (b) estimate parameters for all terms in the model. These assumptions are similar to assuming that interactions are zero in a Latin square design. The estimated mean squares output will tell the researcher which interaction must be assumed to be zero in order to eliminate confounding and make estimates for all terms. MULTIVARIANCE will also identify confounded effects in the case of incomplete block designs.

The expected mean square output may also identify confounding where the researcher does not expect it. For example, adding covariates to a balanced analysis of variance design will produce confounded expected mean squares for terms in the model, and the researcher may wish to adjust for this confounding.

Unbalanced Split-Plots. Both MAD/RUMMAGE and MULTIVARIANCE are the only programs known to the author that can handle unbalanced split-plot designs. Unfortunately, a problem arises with both programs in analyzing these designs: When the model statement for the analysis is written in the customary manner, the core storage required by these designs becomes extremely large, exceeding the capacity of nearly all computer installations. Only designs based on very small sample sizes can be processed in the customary manner. A procedure for getting around the core space problem with the MAD/RUMMAGE program is given later in this paper. Future updates of the RUMMAGE program will probably incorporate this procedure as an automatic part of the output for split-plot designs. This would be desirable, since it is impractical

to obtain multivariate tests for split-plot designs with large sample sizes as the program is written now. MULTIVARIANCE provides a method of getting around the core space problem by transforming the raw data, calling for multivariate tests, and then picking up selected statistics from the multivariate output. The details of this statistical procedure have been described by Boek (1975). A second run is required to get the correct means, since a transformation was made on the raw data in the initial run. One problem with this procedure is that the output for the split-plot designs obtained in this way is not labeled correctly. A separate run can be made on another program like MAD/RUMMAGE or BMD to correctly identify statistics in the MULTIVARIANCE output. MULTIVARIANCE is the only program that can currently perform multivariate tests for unbalanced, split-plot designs that are based on large sample sizes.

Data Management. The model statement for MAD/RUMMAGE is easy to write. The control card command structure for RUMMAGE has been greatly simplified from the original MAD version of the program. However, the program (a) is poorly documented, (b) has rigid requirements for the form that the input data must be in to be accepted by the program, and (c) has no missing data option. The independent variables must be numbered consecutively from one to the number of levels of the variable and must be in sorted order. These requirements mean that the user with a large data file must enter a program like SPSS to recode variables, if necessary, and eliminate missing data, pass this data on a temporary scratch file to a utility program that can sort the independent variables, and then pass this file to the final job step where the analysis is made by MAD/RUMMAGE. This can be accomplished in one run but is inconvenient for the user. MULTIVARIANCE provides a user supplied subroutine for missing data, and concatenation with SPSS for data selection, recoding, etc. MULTIVARIANCE provides a variety of options for inputting the data into the program. In general, MULTIVARIANCE is considerably more convenient than MAD/RUMMAGE for the user with large data files making multiple analyses with the same or similar designs. The MAD/RUMMAGE program provides useful information about confounding and about hypothesis testing with random and mixed models.

Discriminant Analyses. MULTIVARIANCE provides a wider variety of multivariate statistics than MAD/RUMMAGE including discriminant analyses and canonical correlation. It is often useful to follow up significant multivariate analyses of variance tests with discriminant analyses to identify the particular dependent variables that were influenced most by a given independent variable. MULTIVARIANCE can provide discriminant analyses for any term in any multivariate analyses of variance design. Discriminant analyses are not available with MAD/RUMMAGE. RUMMAGE will, however, provide analyses of categorical data, as described in the Categorical Data section of this paper for log-linear models.

A METHOD FOR ANALYSIS OF SPLIT-PLOTS

One of the chief limitations of the MAD/RUMMAGE Program, as it is currently written, is its inability to process split-plot designs based on a large sample size. Even moderately sized samples very quickly exceed the core limitations of most computer centers. This problem is unique to split-plot designs. Other designs, like factorial designs, can readily be processed even with very large sample sizes. A procedure for getting around the core space problem with MAD/RUMMAGE is given in this section of the paper.

The problem arises with split-plot designs because they have more than one "error" term. These designs include one whole-plot error term that is used to test the significance of between subjects or plots (nonrepeated-measures) terms, and one or more split-plot error terms that are used to test the significance of the repeated-measures or split-plot term(s) and interactions with these term(s). The whole-plot error term consists of a random subjects or plots term nested within the between subjects or plots (nonrepeated) terms, while the split-plot error term(s) consists of the interactions between each repeated-measures (split-plot) term and the whole-plot error term. The model statement in the current MAD/RUMMAGE program allows a person to include any number of "error" terms in the model statement; however, only the last of these error terms does not add to the core space required by the computer. Each error term except the last one adds a dramatic amount to the required core space.

A method for analyzing these split-plot designs, suggested by Hendrix (1975), involves dividing the problem between the MAD/RUMMAGE program and another program that can handle balanced repeated-measures designs. This approach has several disadvantages: (a) It requires writing model statements that are unique to the dividing procedure; (b) it requires a fair amount of hand calculation (subtraction); (c) it requires two computer programs; and (d) it cannot handle multivariate analysis of variance.

Split-Plot Example

A different method for analyzing these split-plot designs, which has some of the previous limitations but can be handled within the MAD/RUMMAGE program alone, is shown below. The complete full-rank model of a split-plot design can be written as follows for the MAD/RUMMAGE program:

$$Y(IJKL) = T(I) + S(J) + TS(IJ) + C(IJK) + R(L) + TR(IL) + SR(JL) + TSR(IJL) + CR(IJKL) + E. \quad (1)$$

In this design, the terms T and S are between subjects or plots (non-repeated) terms and R is the repeated-measures or split-plot term. As the model is written above, C is the whole-plot error term and is used to test the significance of the terms which precede it in the model and CR is the split-plot error term. The term E, as written above, contains no degrees of freedom and serves only to terminate the model. The C and CR terms above require a great deal of core storage. The MAD/RUMMAGE program is written so that the E term terminates the model and also collects the sums of squares due to any terms that are deleted from the complete full-rank model. This being the case, it is possible to immediately delete the CR term from the model (as it is written above) and let the sums of squares for this term be collected by the E term. However, the C term will still make the problem exceed storage capacity for all but the smallest samples. Both the C and the CR terms can be deleted from the model as follows:

$$Y(IJL) = T(I) + S(J) + TS(IJ) + R(L) + TR(IL) + SR(JL) + TSR(IJL) + E. \quad (2)$$

In this case, the E term contains the sums of squares for both the C and CR terms. The sums of squares, degrees of freedom, estimated means, etc., for all other terms in the model besides E are correct. The problem now becomes one of separating the sums of squares for the C and CR terms that are confounded within the E term.

A separate run can be made on MAD to obtain the correct sums of squares for the C term, and then the correct sums of squares for the CR term can be obtained by subtraction from the E term listed in (2) above. To be specific, the individual responses or scores can be summed across all levels of the repeated factor R, and these sums can be run with a MAD/RUMMAGE model that includes the between-subjects or plots (nonrepeated) terms, T and S, and excludes the repeated-measures term R:

$$Y(IJ) = T(I) + S(J) + TS(IJ) + E. \quad (3)$$

The sums of squares for the E term in (3) above are equivalent to the sums of squares for the C (whole-plot error) term in (1), after the E term in (3) has been divided by the number of levels of the repeated factor R. The sums of squares for the whole-plot error (nonrepeated) error term are thus obtained by dividing the E in (3) by the number of levels of the repeated factor R, and the sums of squares for CR are obtained by subtracting the whole-plot error term from the E in (2). The number of degrees of freedom for the whole-plot error term as obtained in (3) is correct, and the number of degrees of freedom for the CR term is obtained by subtracting the number of degrees of freedom for the whole-plot error term from the number of degrees of freedom given for the E term in (2).

General Split-Plot Procedure

The above approach can be generalized to any split-plot analysis of variance design as follows:

1. Any split-plot design can be analyzed, but each design will require as many separate runs with different model statements as there are error terms in the model.

2. The first run should include all terms in the complete full-rank model except for the error terms, which should be deleted. This run will produce the correct sums of squares and degrees of freedom for all terms included except for the E term, which will contain the sum of the sums of squares and degrees of freedom for all error terms in the model.

3. The whole-plot error term can be obtained in a separate run by summing individual scores across all levels of each repeated-measures factor in the model. If there is more than one repeated-measures factor, these scores should be summed over all levels of all of these factors. These sums are then run on MAD/RUMMAGE using a model statement that includes the terms tested by the whole-plot error and excludes the terms tested by the split-plot error(s). The sums of squares for the E term of this model is divided by the sum of the number of levels of the repeated-measures factor(s) in the model.

4. When there is more than one split-plot error term, one run with a distinct model statement is required for each split-plot error term in the model, except for the one that is entered last.

a. The first split-plot error term is obtained by summing individual scores across all levels of the repeated-measures factor(s) except for the repeated-measures factor that enters into the error term being obtained. A hypothetical repeated measures factor B, for example, should be tested by the B x subjects interaction, so in this case individual scores should be summed across all levels of repeated-measures factors that happen to be in the model except for B. A run is than made on the MAD/RUMMAGE with a model that includes all terms tested by the whole-plot error and all terms tested by the B x subjects interaction. All terms tested by all other split-plot error terms that are in the model are excluded. All error terms except for the final E should, of course, also be excluded from the model statement, so that in this case the E collects the sums of squares for the whole-plot error plus the B x subjects interaction. The sums of squares for the E term resulting from this run should be divided by the sum of the levels of repeated-measures factors that are in the model <u>besides</u> the B factor. The correct sums of squares for the B x subjects interaction can be obtained by subtracting the whole-plot error term from the E term obtained in this run that has been divided by the number of levels of repeated factors as given above.

b. The second split-plot error is obtained in the same manner as the first, by (a) summing individual scores across levels of the repeated factors that do not enter into the error term of interest, (b) analyzing these sums with a MAD/RUMMAGE model that includes the terms tested by the whole-plot error as well as the terms tested by the split-plot error of interest, (c) dividing the resultant sums of squares for the E term by the number of levels of repeated factors that went into the initial sums, and (d) taking this result and subtracting the whole-plot error from it.

c. The final split-plot error term can be obtained by subtracting each of the previously obtained error terms from the E term that was obtained in the first run. Degrees of freedom for this final split-plot term should also be obtained by subtracting the degrees of freedom for previously obtained error terms from the degrees of freedom for the E term obtained in the first run. The E term from this first run collected the sums of squares and degrees of freedom for all error terms.

Limitations

The preceding approach will only work when balance exists across all levels of each of the repeated-measures factors. In other words, there needs to be one observation per cell across all levels of each repeated-measures factor. If missing data exist at one level of a repeated factor, the data for all levels of the repeated factor need to be removed. However, the previous approach is appropriate when unbalance (unequal cell sizes) exists for the nonrepeated, between-subjects factors.

A one-way repeated measures design cannot be divided into two runs on the MAD/RUMMAGE program. The "whole-plot" error cannot be obtained with a separate run, since there is no term in the model tested by "whole-plot" error. However, the initial MAD/RUMMAGE run can be made, deleting the random subjects factor that is used as a blocking factor. The error term will then include the subjects x repeated-factor error term plus the random subjects factor. The random subjects factor could be calculated with a separate Fortran routine. However, one-way repeated measures designs can readily be handled with other programs in the univariate case. Unfortunately, most other programs cannot handle the multivariate case, and the core space problem may prohibit the multivariate case from being run with the MAD/RUMMAGE program.

The preceding approach is not practical for multivariate analysis of variance with any split-plot design. With multivariate analysis of variance there is a sums of squares and cross products matrix associated with each term in the model. The test statistic in multivariate analysis of variance is the determinant of the error matrix divided by the determinant of the sum of the error matrix plus the matrix for the term

being tested. The determinants are in error when using the preceding subtraction approach with multivariate analysis of variance, since the determinant of the difference between matrices is not equal to the difference between the determinants of two matrices. To obtain the correct determinants, the matrix for the whole-plot error would have to be subtracted from the matrices for the other terms in the model and determinants calculated for these differences. Although the appropriate matrices can be obtained from the MAD/RUMMAGE package, the amount of calculation required to subtract matrices and calculate determinants obviously exceeds what is practical to do by hand.

CATEGORICAL DATA

Psychologists collect data that are generally measured on nominal or ordinal scales. Results are often expressed in the form of frequency tabulations in one-way, two-way, and multiway tables. Such data are often analyzed with the traditional Pearson chi-square statistic as applied repeatedly to different subsets of the total possible number of two-dimensional tables. Army researchers often have occasion to measure such nominal variables as race, sex, MOS, mission type (combat, support), etc. Categorical data of this nature are frequently analyzed by the repeated application of the Pearson chi-square statistic to all possible combinations of two-way tables, using the SPSS Crosstabs procedure. Even ordinal data, including ordinal questionnaire responses, are often expressed in terms of the percentages of subjects who selected particular responses, particularly since data presented in this way are easily interpreted by nonresearchers in terms of the original scales. However, it is often difficult, and in some cases impossible, for a researcher to test the hypotheses of interest in terms of two-way contingency tables. In many cases the researcher runs multiple tests in order to test hypotheses that have been stated in a fragmentary form.

The use of linear regression models for the analyses of multidimensional categorical tables has been described by Grizzle, Starmer, and Koch (1969). This approach provides a comprehensive method for the statistical analysis of qualitative data that is directly analogous in scope and power to multiple regression and multivariate analysis of variance as applied to quantitative data (Koch & Reinfurt, 1970). This approach provides a better method for testing many hypotheses than the repeated application of the Pearson chi-square. Applications of this methodology are beginning to appear in the social science literature (see Giles, Gatlin, & Cataldo, 1976). This least squares approach to the analysis of categorical data has been programed and is available as a Fortran program called GENCAT (Landis, Stanish, Freeman, & Koch, 1976).[2]

---

[2]This program has been implemented at IBM 360/370 installations. It will shortly be modified to be compatible with non-IBM machines. The program is available from Dr. Richard Landis, Dept. of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109. Army Research Institute, Presidio of Monterey Field Unit, has a copy of this program.

Defining Categorical Data Models

Just as with analysis of variance, the type of analysis that is made with categorical data depends on the specifications of the underlying model. The underlying model depends on the sampling plan of the experiment. The first step in specifying the model for the analysis is to distinguish between (a) the variables that measure the experimental conditions or subgroups to which subjects belong and (b) the variables that measure what subsequently happens to subjects. All possible combinations of levels of the variables measuring experimental conditions or subgroups define the "populations" or "factors," in the design and the possible combinations of variables measuring what happens to subjects define the "responses." A table of proportions defined by the number of response combinations, by the number of populations, is entered into GENCAT.

Multidimensional contingency tables are entered into GENCAT according to how the model has been defined in terms of populations and responses. Several general types of models can be identified:

1. No factor, multiresponse;
2. Unifactor, multiresponse;
3. Multifactor, uniresponse; and
4. Multifactor, multiresponse.

Only Models 1 and 2 can occur with two-way tables, and Models 1 through 3 for three-way tables; otherwise all models can occur.

Model 1. The questions that are asked in the case of the no factor, multiresponse model are analogous to questions that would be asked in repeated-measures analysis of variance designs where all the factors (one or more) in the design are repeated-measures factors. A problem of interobserver agreement could also fit under Model 1. Since all observers rate the same person/situation, the ratings fit in the mold of a repeated-measures analysis of variance design. However, in this case hypotheses of interest would include not only tests of the differences between proportions but also agreement hypotheses: Is agreement different from that expected by chance alone?

Model 2. In the case of unifactor, multiresponse tables, the questions asked are analogous to those asked with one-way multivariate analysis of variance designs. In designs of this nature, the researcher is interested in the association among dependent or response variables as well as the influence of the independent or factor variable on the response variables. With one factor and a series of r response categories, questions asked include (a) the influence of the factor on the marginal distribution of the response, and (b) the influence of the factor on the joint distribution of the r response categories.

Model 3. In the case of multifactor, uniresponse tables, the design is directly analogous to factorial analysis of variance designs. Here the researcher wishes to determine how the factors or independent variables combine to produce the response or dependent variable. The researcher can test for "main effects" for the factors and for "interaction effects," just as in analysis of variance, except in this case the researcher is looking at differences between proportions instead of means. An example of multifactor, uniresponse problems, may be instructive at this point. Table 1 presents a hypothetical factor by response matrix of proportions in a form that could be entered into GENCAT. Both columns of proportions are entered into the program, but a transformation matrix is entered to eliminate the second column, because (a) we are only interested in comparing proportions who received Article 15's, and (b) computations cannot be made when singularity exists (i.e., when the rows add up to 1.0). Singularity also exists when a proportion in the table is zero. When a zero enters into the table, the levels of the factors must either be collapsed to eliminate the zero, or else the zero must be replaced by a small proportion to eliminate the singularity. The GENCAT output for Table 1 would include one chi-square statistic testing significance for the main effect of Race, one for Rank, and one for the Race x Rank interaction.

Table 1

Example of Multifactor, Uniresponse Problem

Race	Rank	Proportion receiving AR-15	Proportion not receiving AR-15
Black	Enlisted	.30	.70
Black	Officer	.00	1.00
White	Enlisted	.20	.80
White	Officer	.02	.98

Note. Race and Rank define the factors or populations; the response is defined by receiving or not receiving Article 15 punishment.

The GENCAT results can be briefly compared to traditional results. Separate Pearson chi-square statistics could have been readily computed in Table 1 for the effects of Race and Rank, but not for the interaction between these factors. With the GENCAT approach, each term in the model is adjusted for the other terms, in a manner analogous to least squares analysis of variance or multiple regression--which would not have been the case had two Pearson chi-square statistics been computed. Also,

when multiresponse models are entered, GENCAT automatically adjusts for the correlation between the multiple responses. The adjustment is made by defining the appropriate variables as "response" instead of "factor" variables. Traditional texts have always noted that the appropriate chi-square test statistic depended on whether the variables were correlated or not (Ferguson, 1966). Fortunately, adjustment for correlation can be made by the appropriate selection of the model for GENCAT.

Model 4. Finally, multifactor, multiresponse models are analogous to factorial, multivariate analysis of variance designs, or to split-plot analyses of variance designs. In these designs, questions are asked about relationships among the dependent or repeated-measures variables as well as the way factors combine to affect the responses.

An Example--The ARI Representation Index

Appropriate test statistics are derived by entering a sequence of transformation design and contrast matrices into GENCAT. These matrices operate on the vector of proportions in such a manner as to define the specific contrasts of interest. Linear, logarithmic ($\log_e$), and exponential transformations of the proportions are possible, and these transformations affect the nature of the hypotheses that are tested. The following ARI research example shows how transformations affect the nature of the hypotheses that are tested.

As one approach to identifying possible areas of institutional discrimination in the Army, Nordlie, Thomas, and Sevilla (1975) constructed a Representation Index[3] as a quantitive measure of how promotions, punishments, education, etc., have been distributed among whites and nonwhites. This Representation Index is numerically equivalent to a simple linear transformation of the ratio of two proportions. In other words, with this index we are comparing the ratio of the proportion of blacks who receive a given action to the proportion of whites who receive this action, and then transforming this quantity linearly so it will have an origin of zero. So far no statistical tests have been made to test whether or not a given Representation Index is significantly different from zero or whether the Representation Index for one group (e.g., blacks) is significantly different from the one for another group (e.g., Spanish). With Army-wide samples, these tests may not always be relevant; however, tests of this nature are important

---

[3] $\text{Representation Index} = \left(\frac{\text{actual number}}{\text{expected number}} \times 100\right) - 100,$

where actual number equals the number of minorities receiving a particular action, and expected number equals the expected proportion if there is no association between the event and skin color, times the number of individuals receiving the particular action.

when representation indexes are computed with smaller sample sizes, and the researcher wants to know if chance variation is responsible for the size of the indexes. A Pearson chi-square statistic could be computed on the difference between proportions, but the Representation Index measures a ratio rather than a difference. To test the significance of the Representation Index (or the ratio) directly, a logarithmic transformation is made on the two proportions, and then the test statistic is computed on the difference between the logarithmically transformed porportions, which is in fact a test of the ratio of the proportions. Tests of significance for the Representation Index can readily be made with GENCAT. These tests of significance can also be made using a categorical data feature of the RUMMAGE program.

As this example demonstrates, transformations affect the nature of the hypothesis that is tested. Transformations are often made with analysis of variance in an attempt to normalize the distribution of scores. It should be noted that these transformations alter the nature of the hypothesis that is tested as well as the nature of the distribution of scores.

Categorical Data Versus Analysis of Variance

There are several advantages to analyzing data using the GENCAT rather than analysis of variance:

1. The GENCAT approach requires the researcher to make fewer assumptions about the nature of the data.

2. Much of the data collected by psychologists is nominal or ordinal, often not very reliable, and can best be represented as categorical variables.

3. Results can be expressed as percentages and, as such, are readily interpretable by the Army and other nonresearchers.

In many cases, however, results may not differ much from analogous analysis of variance results. Also, the documentation for the program is written by statisticians writing to an audience well versed in matrix manipulations. The way in which transformation, design, and contrast matrices are entered into GENCAT is not at all obvious to nonstatisticians who have not worked with the program.

CONCLUSION

Each of the generalized programs mentioned previously--MAD/RUMMAGE, MULTIVARIANCE, GENCAT--are large programs that took several man-years to write (e.g., MAD has over 12,000 Fortran commands). Bugs have been eliminated over several years' experience with the programs. Together

these programs provide powerful tools for psychologists doing field research. Psychologists often find themselves with (a) unequal sample sizes in field experiments due to lack of control, (b) multiple dependent variables as part of an evaluation research design, and (c) large quantities of nominal or ordinal categorical data. The generalized software packages described here can handle many of the analysis requirements for the types of data listed above.

REFERENCES

Bock, D. Multivariate Statistical Analysis in Behavioral Research. New York: McGraw-Hill, 1975.

Bryce, G. R., & Carter, M. W. MAD--The Analysis of Variance in Unbalanced Designs--A Software Package. Compstat 1974, Proceedings in Computational Statistics. ed: Bruckmann, G., Ferschl, F., & Schmetterer, L. (Eds.). Vienna, Austria: Physica Verlag, ISBN 3 7908 0148 8, 1974.

Dixon, W. J. (Ed.). BMD: Biomedical Computer Programs, 3rd Ed. Berkeley, Calif.: University of California Press, 1973.

Ferguson, G. A. Statistical Analysis in Psychology and Education. New York: McGraw-Hill, 1966.

Finn, J. A General Model for Multivariate Analysis. New York: Holt, Rinehart & Winston, 1974.

Giles, M. W., Gatlin, D. S., & Cataldo, E. F. Racial and Class Prejudice: Their Relative Effects on Protest Against School Desegregation. American Sociological Review, 1976, 41, 280-288.

Grizzle, J. E., Starmer, C. F., & Koch, G. G. Analysis of Categorical Data by Linear Models. Biometrics, 1969, 25, 489-504.

Harris, R. J. A Primer of Multivariate Statistics. New York: Academic Press, 1975.

Hendrix, L. Suggestions on Analyzing Split-Plots. Provo, Utah: Brigham Young University, Statistics Department, February 1975.

Koch, G. G., & Reinfurt, D. W. The Analysis of Complex Contingency Table Data from General Experimental Designs and Sample Surveys. (North Carolina Institute of Statistics Mimeo Series No. 716.) Chapel Hill, N.C.: Statistics Department, July 1970.

Landis, R. J., Stanish, W. M., Freeman, J. L., & Koch, G. G. A Computer Program for the Generalized Chi-Square Analysis of Categorical Data Using Weighted Least Squares (GENCAT). (Biostatistics Technical Report No. 8.) Ann Arbor, Michigan: University of Michigan, Department of Biostatistics, April 1976.

Nie, N. H., Hull, C. H., Jenkins, J. G., Steinbrenner, K., & Bent, D. H. SPSS, 2nd ed. New York: McGraw-Hill, 1975.

Nordlie, P. G., Thomas, J. A., & Sevilla, E. R. Measuring Changes in Institutional Racial Discrimination in the Army. ARI Technical Paper 270, December 1975.